EDA CASE STUDY

Group Assignment

Members: Sachin Katiyar and Asawari Kadam

Problem Statement

The loan providing companies find it hard to give loans to the people due to their insufficient or non-existent credit history. When the company receives a loan application, the company has to decide for loan approval based on the applicant's profile. Two types of risks are associated with the bank's decision:

- If the applicant is likely to repay the loan, then not approving the loan results in a loss of business to the company.
- If the applicant is not likely to repay the loan, i.e. he/she is likely to default, then approving the loan may lead to a financial loss for the company.

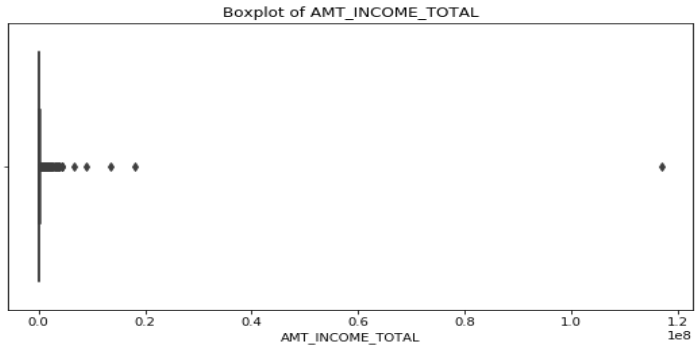
This case study aims to identify patterns which indicate if a client has difficulty paying their installments which may be used for taking actions such as denying the loan, reducing the amount of loan, lending (to risky applicants) at a higher interest rate, etc. In other words, **the company wants to understand the driving factors (or driver variables) behind loan default, i.e. the variables which are strong indicators of default**

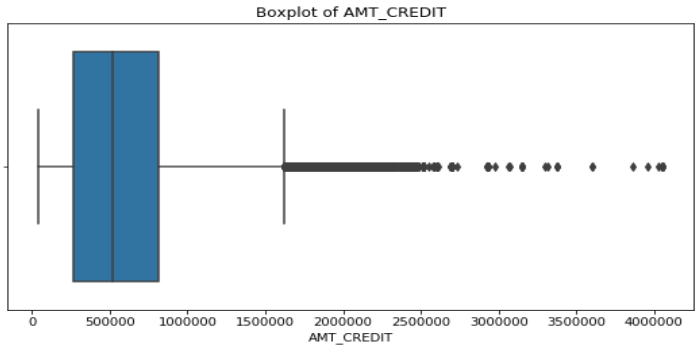
Steps

- Imported data ‘application_data.csv’.
- Checked for missing values and their percentage of missing values.
- Distributed the data into Target 1 and Target 0. Then we filled values for the respective columns.
- On checking columns we see outliers in the dataset.
- We decided to remove outliers for 2 columns, namely AMT_INCOME_TOTAL and AMT_CREDIT.

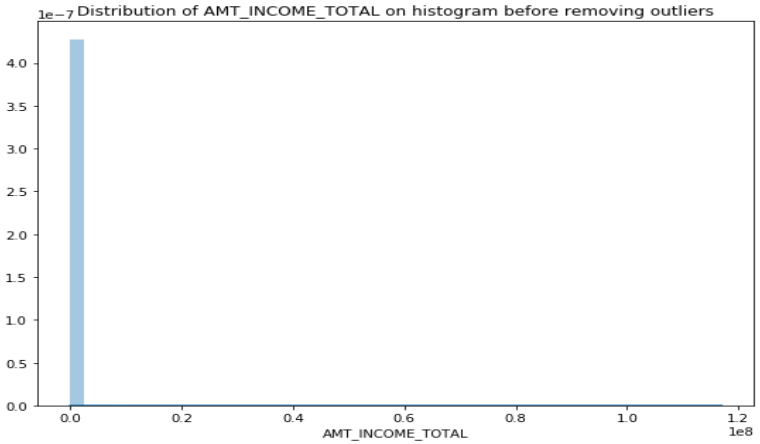
Outlier Analysis

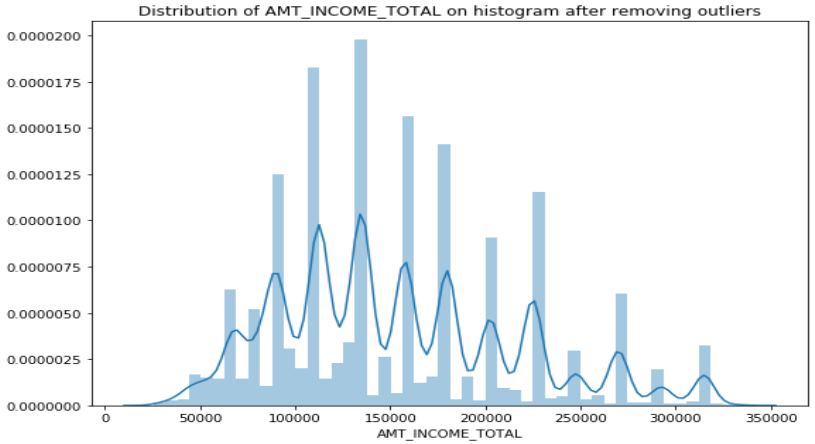
- Before the outlier analysis was done the box plot for both the column :





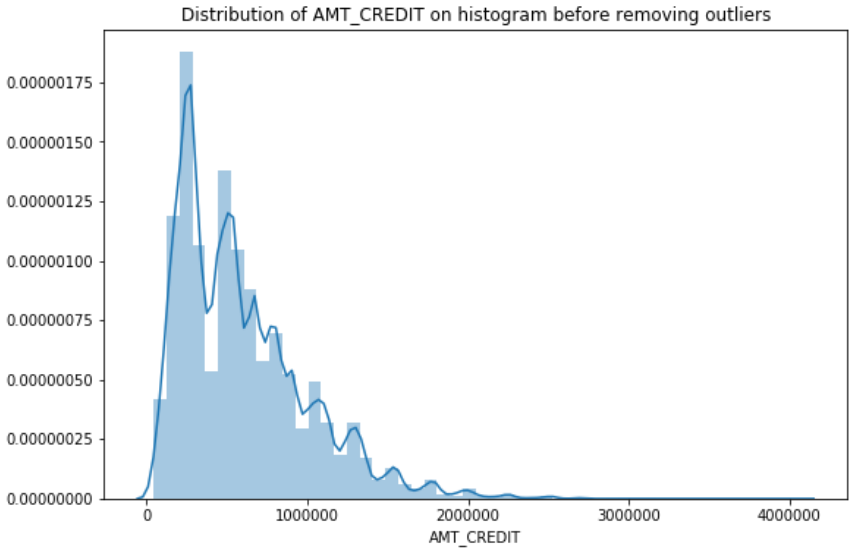
- After the outlier analysis was done and the distribution was as follows for AMT_INCOME_TOTAL:

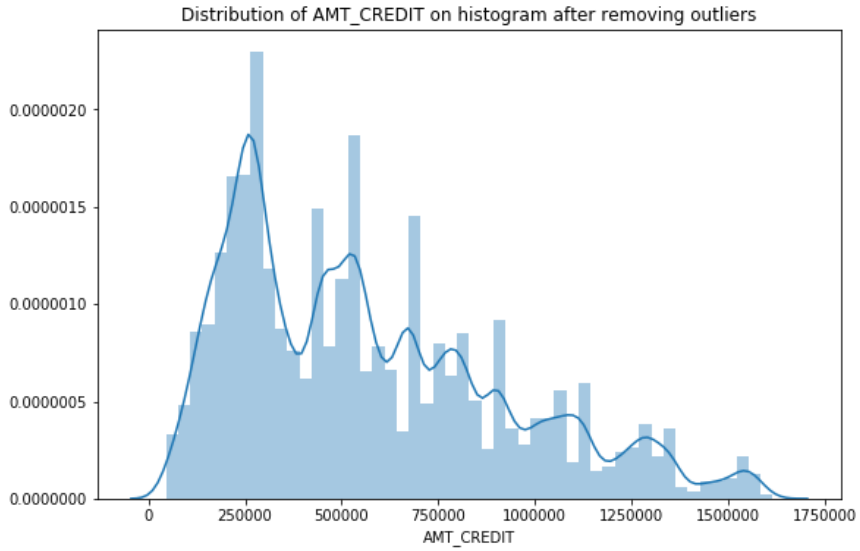




Outlier analysis

- The outlier analysis and distribution for AMT_CREDIT

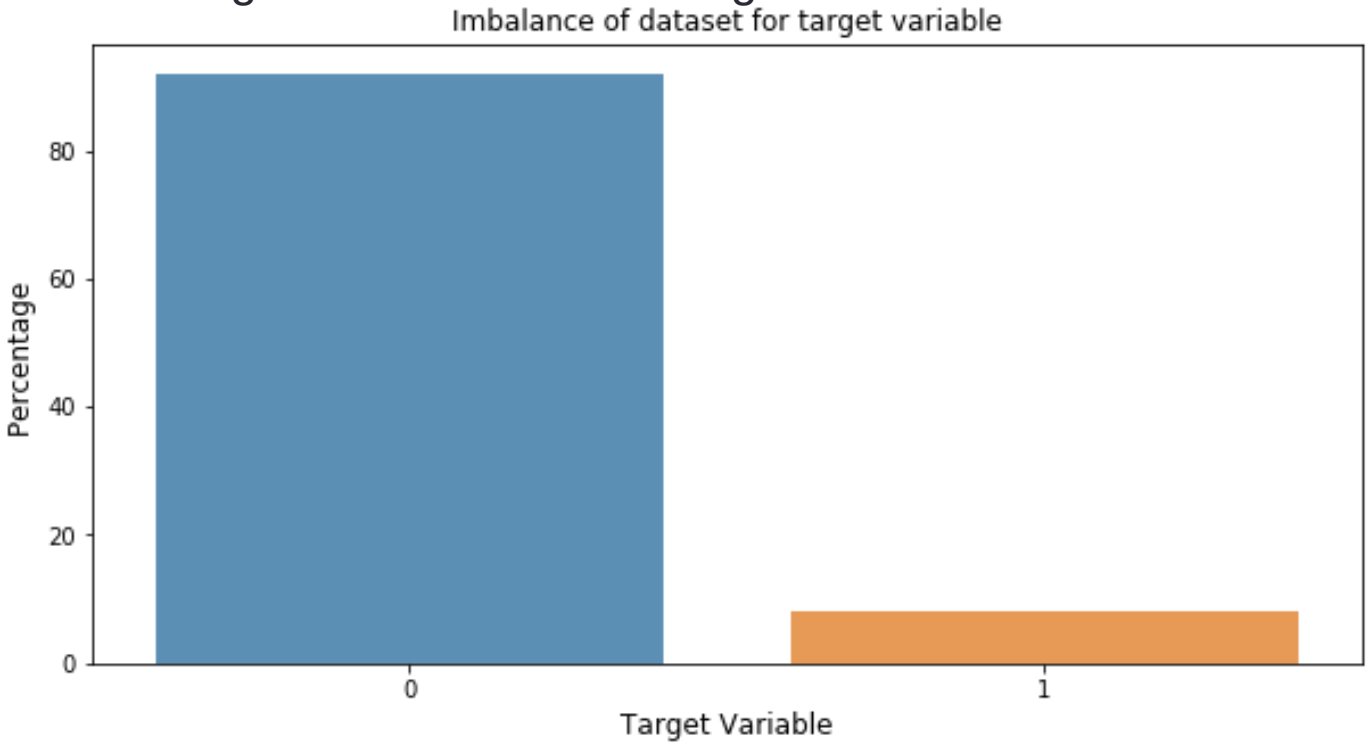




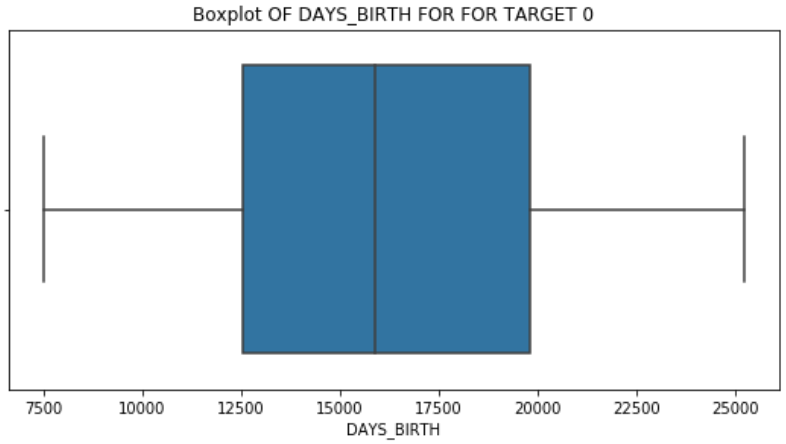
- In both the graphs after the distribution we see a uniformity of data spread across the rows.

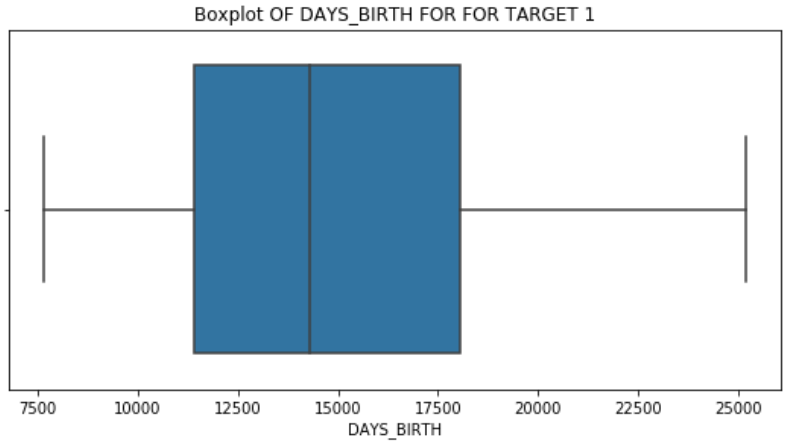
Data Imbalance

- We need to check if the Target data is imbalanced or no. To do this we plot a bar plot and analyze the plot. The following shows an imbalance in data.
- The Ratio of Target variable 1 to the Target variable 0 is: 11.387



Univariate Analysis for Columns

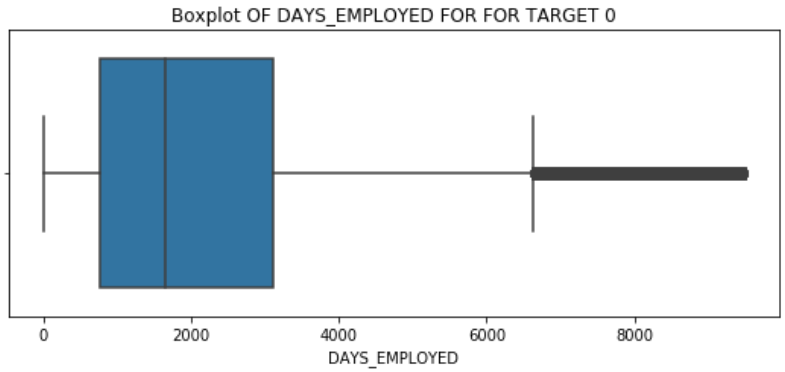


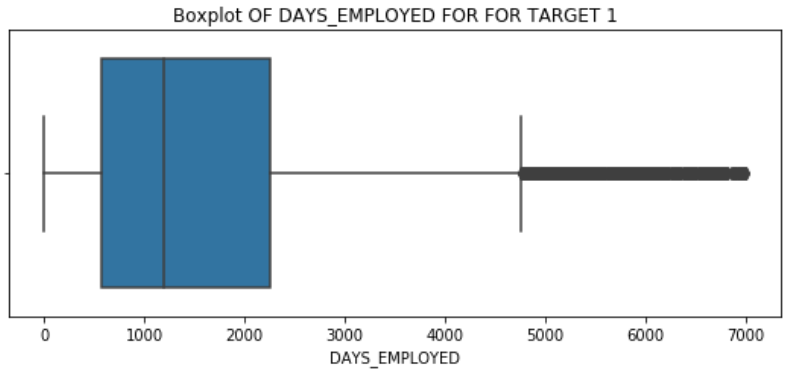


Analysis and Observation:

For Days of birth in Target 0 the 25th to 75th range between 20000 and 12500 whereas in Target 1 it shifts to 18000 around and 11000 around, which means that the people in this range are young people which would not have sufficient balance to repay the loan. So they are more likely to have difficulties in paying the amount . This may be a deciding factor for non-payment of loans

Univariate Analysis for Columns



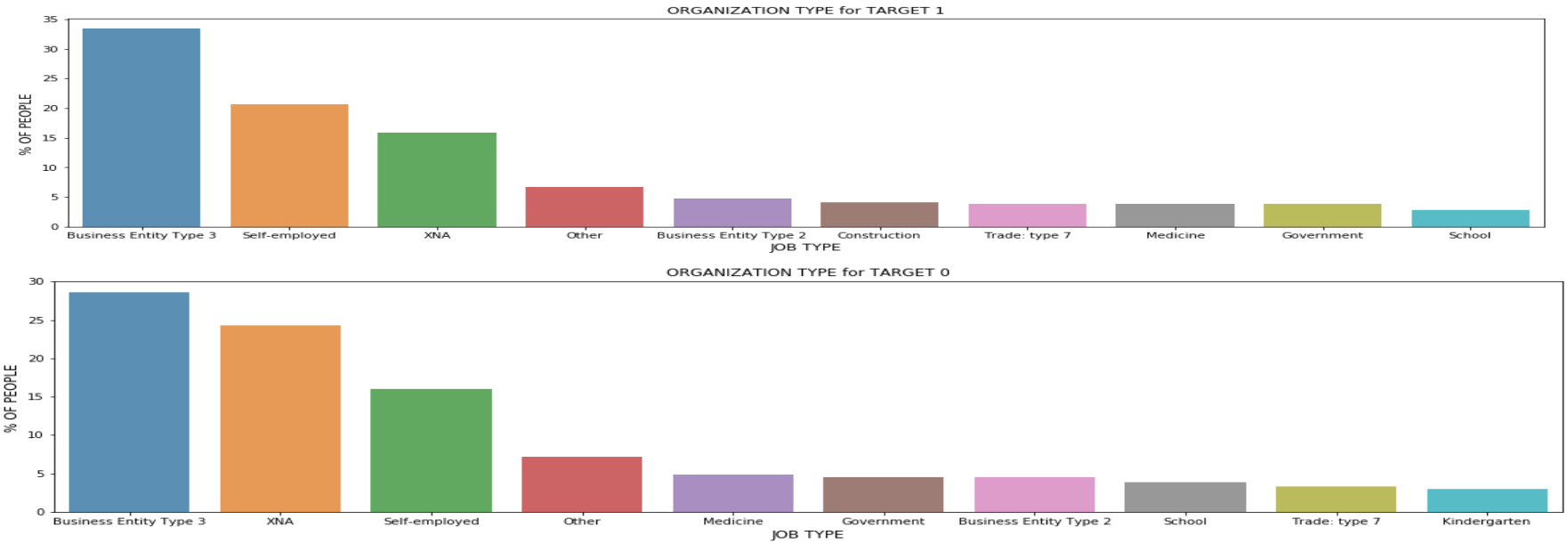


Analysis and Observation:

For Target 1 and 0 we a lot of values are on the outside, meaning we have outliers in this column. We have removed outliers and the box plot is plotted.

For Target 0 we see that the DAYS_EMPLOYED is having its mean around 2000 days whereas for Target 1, DAYS_EMPLOYED mean is around the 1000-1500. This means that the people are fairly new to the organization and have just started their the career. Enough amount of saving to repay the loan would not be possible.

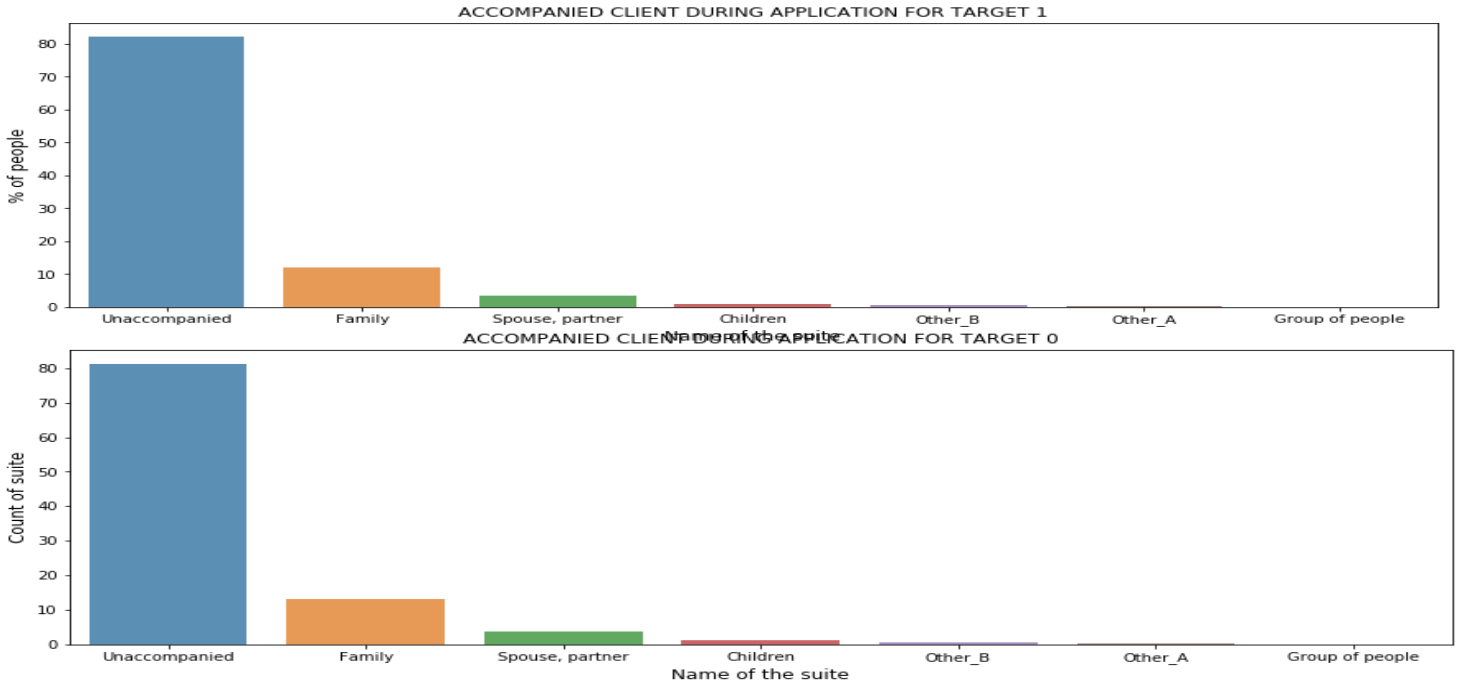
Univariate Analysis for Columns



Analysis and Observation:

For Target 1 we see column Self Employed as the 2nd Highest value contributing to the target 1 whereas in Target 0 it is seen one position lower indicating that the value in target 0 is lesser and having less influence than Target 1. This implies that the people who are self-employed may have a difficulty in payment. Similar are the cases for the other bar plots in comparison with Target 1 and Target 0.

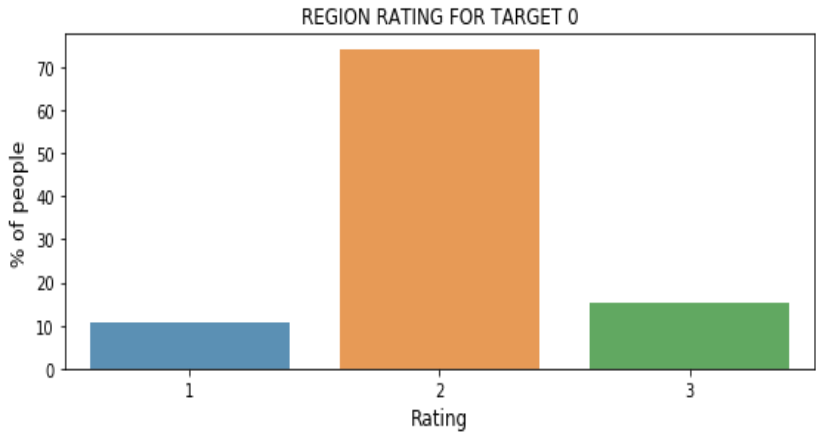
Univariate Analysis for Columns

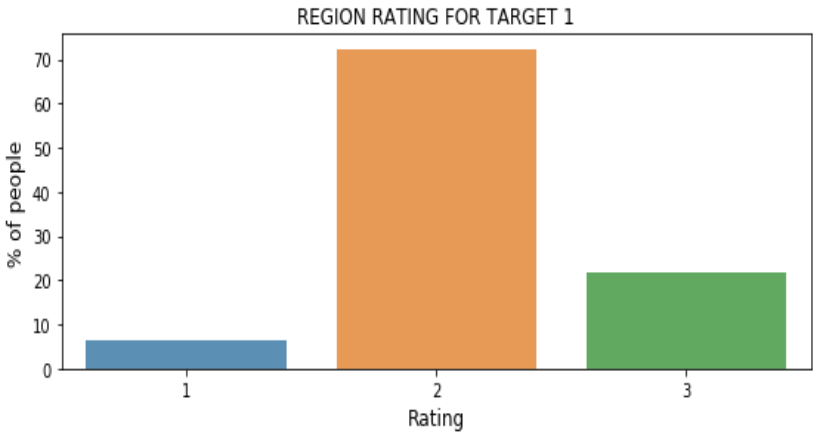


Analysis and Observation:

For the above we see that the maximum number of people who went to apply for the loan were unaccompanied and for both Target 1 and Target 0 the plots are same. Indicating that if the person is unaccompanied or accompanied by someone he is not likely to repay the amount.

Univariate Analysis for Columns

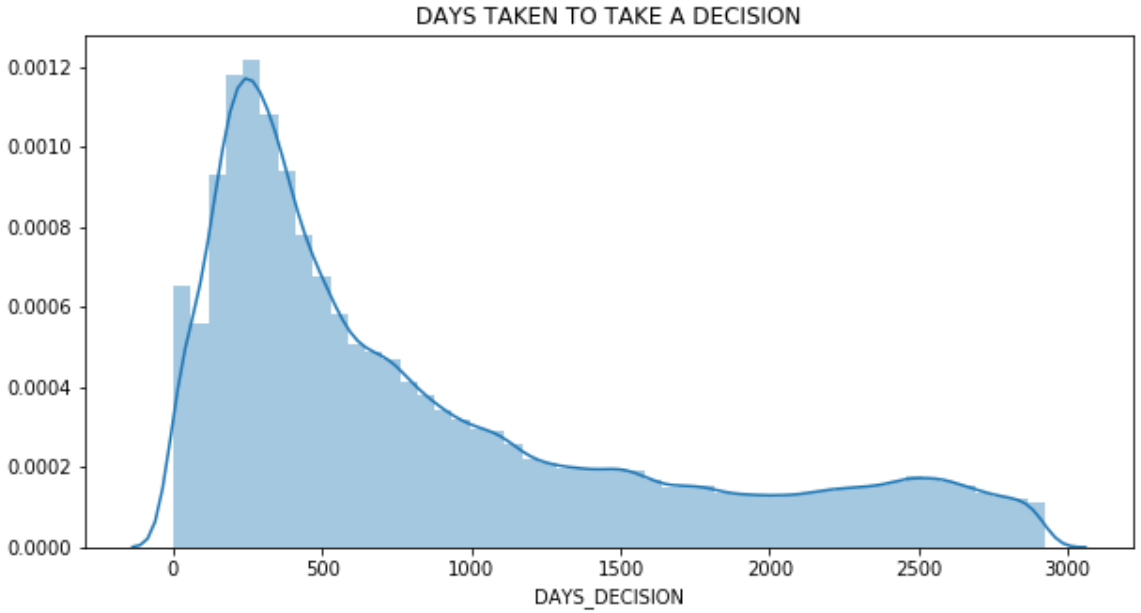




Analysis and Observation:

We see that the Region Rating-2 for both Target 1 and Target 0 are the highest among the 3. Whereas the other 2 values differ slightly in the values. For Region Rating-3, the percent of rating in Target 0 is lying between 10-20 % whereas in Target 1 it increases and lies in the range of 20-30%. Which means the people in Region 3 are likely to have payment difficulties and would not be able to pay the amount than the people of Target 0.

Univariate Analysis for Columns

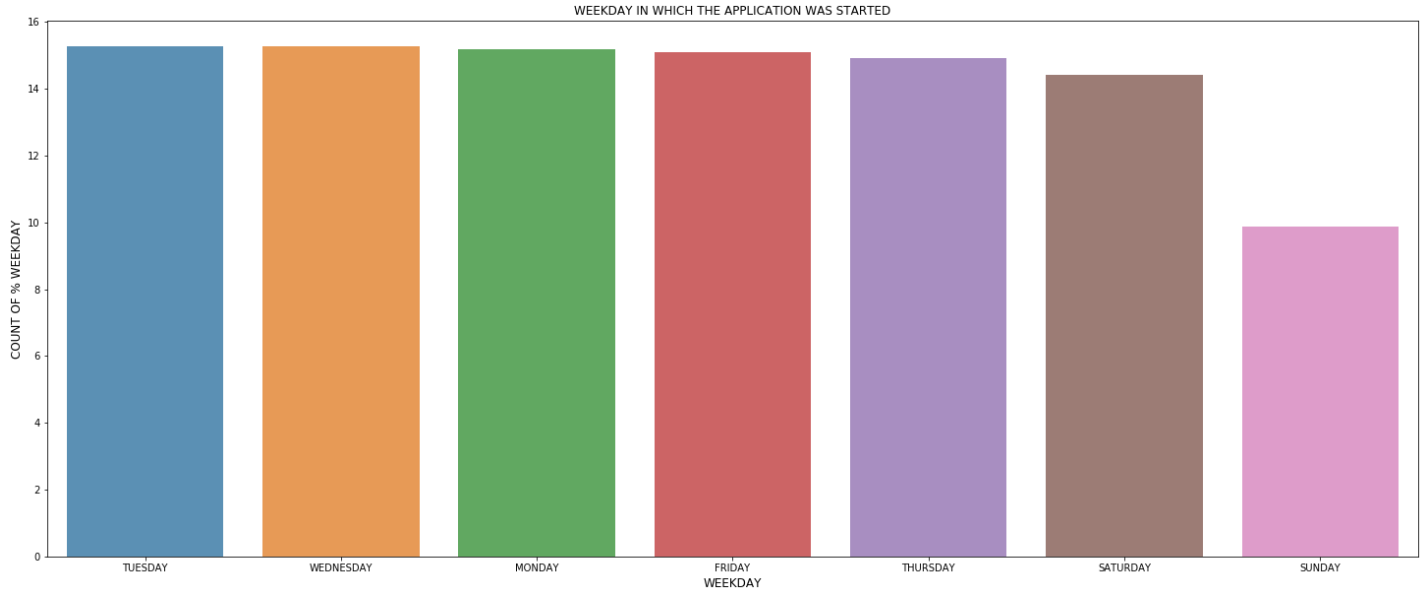


Analysis and Observation:

It is seen that the days to take decision is highest when the previous application is submitted and is approved with 500 days of the current application.

There are very few applicants whose loans approvals is done within a few days of submitting the loan.

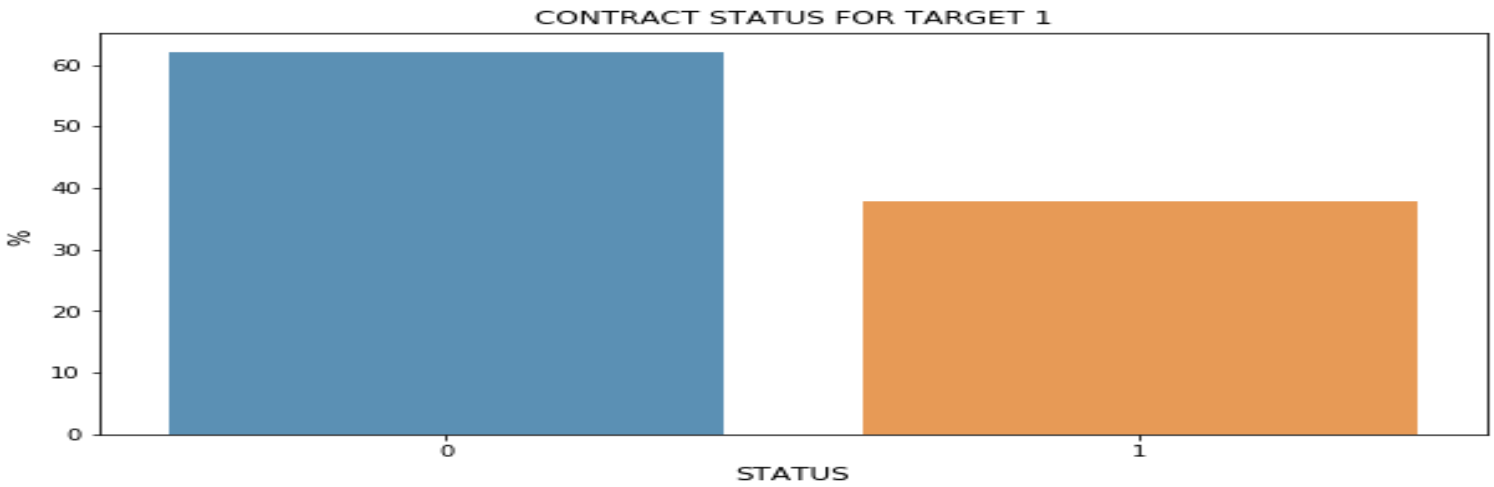
Univariate Analysis for Columns



Analysis and Observation:

This shows us that application work was started on all weekdays equally and works slightly less on Saturday and Sunday

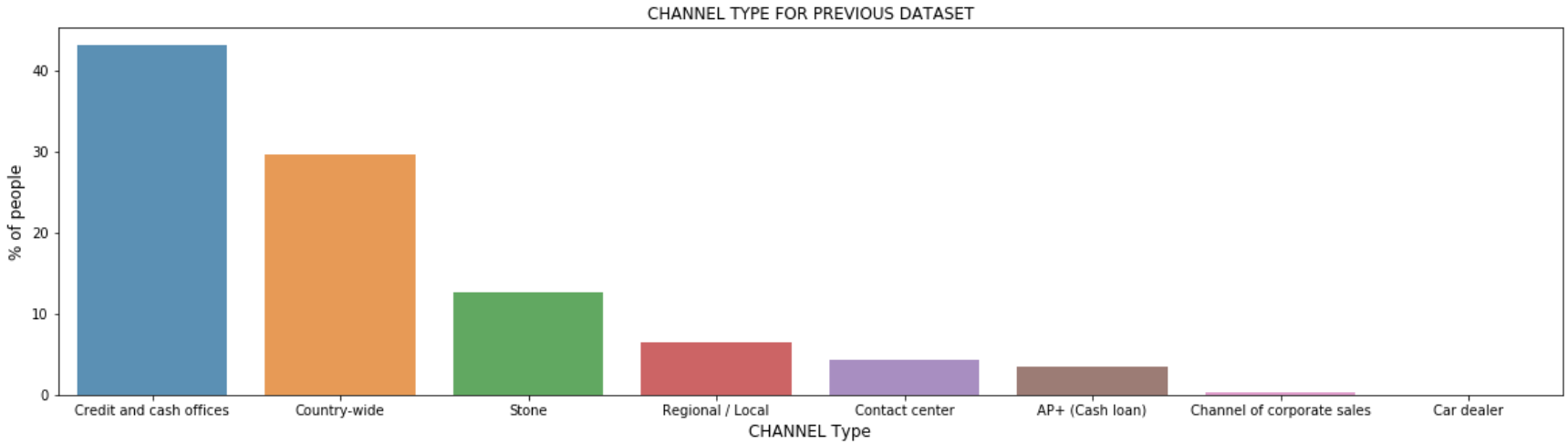
Univariate Analysis for Columns



Analysis and Observation:

On working the column of CONTRACT _STATUS, we combined the rejected,refused, canceled and unused offer to 1 and approved to 0. We see that in previous data file, a 60% of applicant were given loan and around 40% were not given loan.

Univariate Analysis for Columns



Analysis and Observation:

This shows us that the highest no of loan acquiring for Previous application was done through Credit and Cash offices and the lowest was done through a Car dealer.

Bivariate Analysis

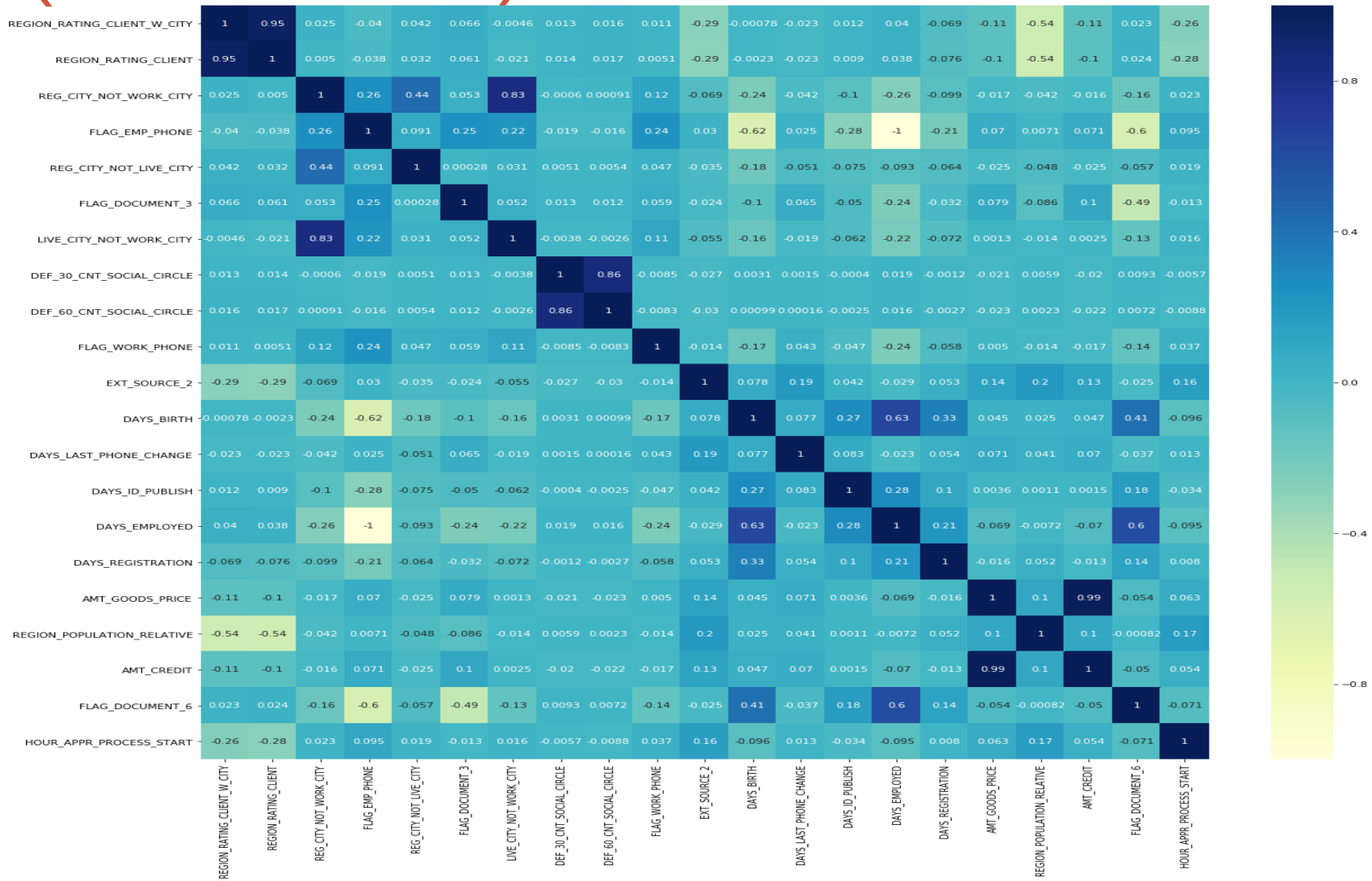
- First we have run correlation on application data to find out columns that affect target variable the most.

```
REGION_RATING_CLIENT_W_CITY   0.061
REGION_RATING_CLIENT          0.059
REG_CITY_NOT_WORK_CITY        0.051
FLAG_EMP_PHONE                0.046
REG_CITY_NOT_LIVE_CITY        0.044
FLAG_DOCUMENT_3               0.044
LIVE_CITY_NOT_WORK_CITY       0.033
DEF_30_CNT_SOCIAL_CIRCLE      0.032
DEF_60_CNT_SOCIAL_CIRCLE      0.031
FLAG_WORK_PHONE               0.029
Name: TARGET, dtype: float64
EXT_SOURCE_2                 -0.160
DAYS_BIRTH                   -0.078
DAYS_LAST_PHONE_CHANGE       -0.055
DAYS_ID_PUBLISH              -0.051
DAYS_EMPLOYED                -0.047
DAYS_REGISTRATION            -0.042
AMT_GOODS_PRICE              -0.040
REGION_POPULATION_RELATIVE   -0.037
AMT_CREDIT                   -0.030
FLAG_DOCUMENT_6              -0.029
HOUR_APPR_PROCESS_START      -0.024
```

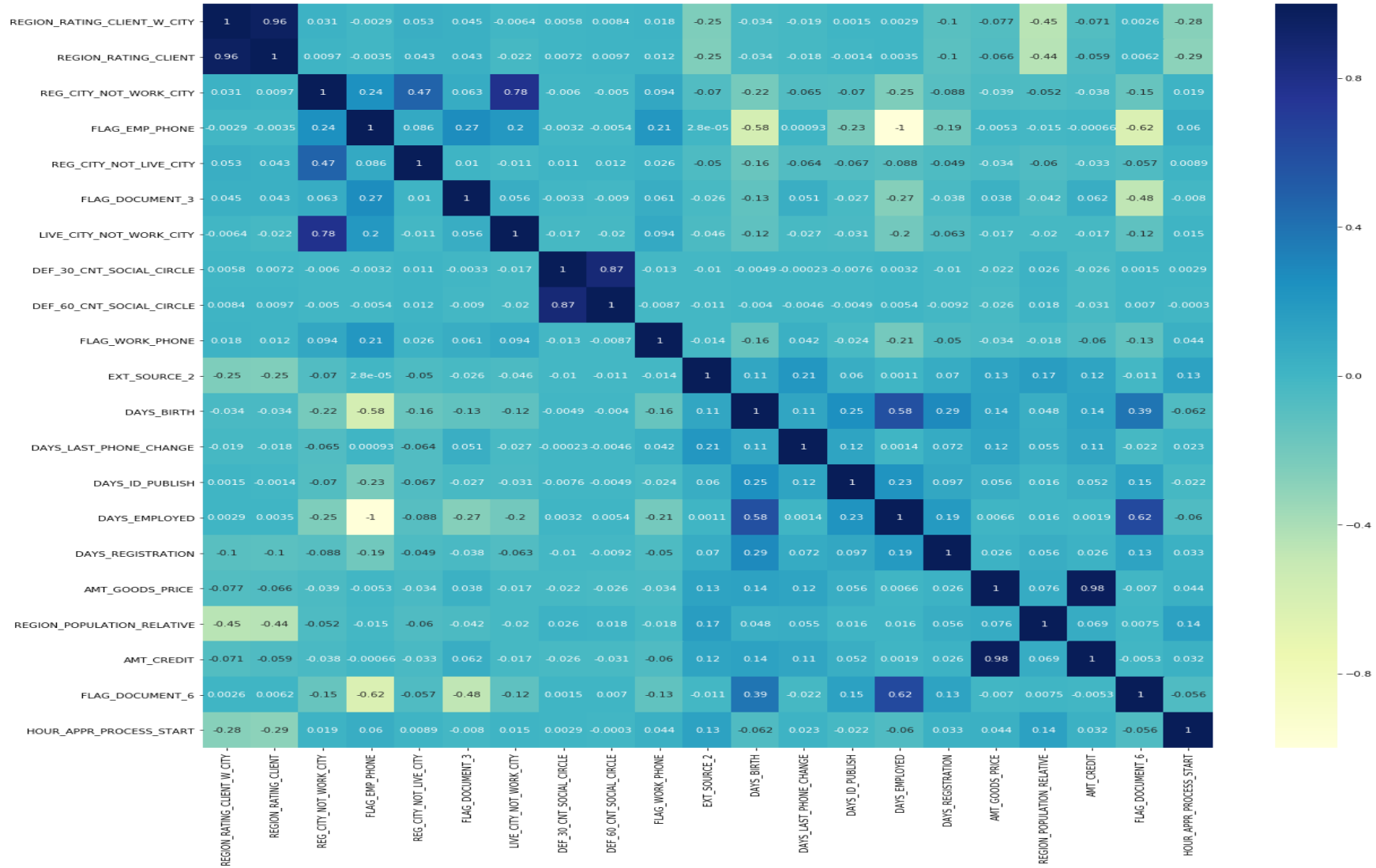
Scenarios based on coorelated columns

- Here columns having positive values are directly proportional and columns having negative values are inversely proportional to the person who will default.
  - If rating of region given by bank is more then it means person coming from that area have more chances of default.
  - If address registered in bank is different from the location where he lives currently then it shows that person can default.
  - If that person has not seen for 30/60 days after past due then he/she has more chance of being defaulter.
  - If days of birth has a smaller value then he/she is a young person and do not have enough money to repay loan and can be default.
  - If person is changing his/her phone frequently then he/she has high chances of becoming a defaulter.
  - If days of employment is less then that person is not financially strong and will not be able to pay loan back in most of the caes.

Correlation matrix for Target variable 0 (will not default)



Correlation matrix for Target variable 1 (having chances of dafault)



Analysis based on both Target Correlation Matrix

- Both matrix are similar mostly.
- Variables that are highly correlated are:
  - REG_CITY_NOT_WORK_CITY – LIVE_CITY_NOT_WORK_CITY
  - DEF_30_CNT_SOCIAL_CIRCLE - DEF_60_CNT_SOCIAL_CIRCLE
  - AMT_CREDIT – AMT_GOODS_PRICE
  - FLAG_DOCUMENT_6 – DAYS_EMPLOYED
  - FLAG_EMPLOYEE_PHONE – DAYS_BIRTH
  - FLAG_DOCUMENT_6 – DAYS_BIRTH

Correlation matrix on Previous Application Dataset